DOCUMENT RESUME

ED 339 752 TM 017 685

AUTHOR Jarrell, Michele G.
TITLE Generating an Empirical Probability Distribution for the Andrews-Pregibon Statistic.
PUB DATE Nov 91
NOTE 14p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (20th, Lexington, KY, November 12-15, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Simulation; Error of Measurement; Matrices; *Multivariate Analysis; *Probability; *Statistical Distributions
IDENTIFIERS *Andrews Pregibon Statistic; Empirical Research; FORTRAN Programing Language; *Outliers

ABSTRACT

A probability distribution was developed for the Andrews-Pregibon (AP) statistic. The statistic, developed by D. F. Andrews and D. Pregibon (1978), identifies multivariate outliers. It is a ratio of the determinant of the data matrix with an observation deleted to the determinant of the entire data matrix. Although the AP statistic has been used many times, no probability distribution has been available for it. The AP statistic is based on the volume of confidence ellipsoids and is a function of leverage and residual. Small values of the AP statistic are associated with outlying observations. A probability distribution was developed through computer generation of 10,438 samples of n=150 and p=3 from a multivariate normal population using a FORTRAN program. The AP statistic was calculated for each observation in each sample. A data file was created with the statistics and sorted by the ratios, and a frequency distribution was run. This distribution was used to obtain the critical values for the various error rates. Changing the parameters in the FORTRAN program could allow development of the probability distribution for other values of n and p, including data that are other than multivariate normal. Eight brief appendixes summarize computer programs that were run on the calculations. One data table and seven references are provided. (SLD)

Generating an Empirical Probability Distribution for the Andrews-Pregibon Statistic

Michele G. Jarrell
The University of Alabama

Paper presented at the annual meeting of the Mid-South Educational Research Association
Lexington, Kentucky, November 13, 1991.

Since 1978, the Andrews-Pregibon statistic has been available for identifying multivariate outliers. The statistic is a ratio of the determinant of the data matrix with an observation deleted to the determinant of the entire data matrix. Although it has been referenced in many statistics books and articles, no probability distribution has been available for the Andrews-Pregibon statistic.

Andrews and Pregibon (1978) suggested a linear model which identified deviant or influential observations by deleting observations, calculating the residual sum of squares, calculating the inverse of the inner product matrix formed after deleting observations, and forming a ratio. The Andrews-Pregibon statistic is based on the volume of confidence ellipsoids (Chatterjee & Hadi, 1988) and is a function of leverage and residual (Fung, 1990). Small values of the Andrews-Pregibon statistic are associated with outlying observations. Wood (1983) found that this procedure solves the problem of masking, which is the inability to detect outlying observations because of the presence of other outliers (Andrews & Pregibon, 1978; Barnett & Lewis, 1978; Hawkins, Bradu, & Kass, 1984), but the number of subsets that need to be examined may be quite large with this procedure.

Andrews and Pregibon (1978) state that this procedure will identify observations which are potential outliers and which are influential on the linear model estimates. The Andrews-Pregibon ratio is expressed as:

$$AP = \frac{\det\ (X_{(i)}^{T} X_{(i)})}{\det\ (X^{T} X)}$$

where **det** is the determinant of the matrix which results from multiplying the two matrices, $\mathbf{X}^T$ is the transpose of the **X** matrix, $\mathbf{X}^T_{(i)}$ is the transpose of the **X** matrix with the *i*th observation deleted, and $\mathbf{X}_{(i)}$ is the **X** matrix with the *i*th observation deleted.

A probability distribution was developed on an IBM 3090\400E mainframe computer using a REXX executable file to run several SAS programs and a FORTRAN program. Over ten thousand (10,438) samples of n = 150 and p = 3 were generated from a multivariate normal (0,1) population using a FORTRAN program developed by Morris (1975). The Andrews-Pregibon statistic was calculated for each observation in each sample. A data file was created with the statistics, the file was sorted by the ratios, and a frequency distribution was run. This distribution was used to obtain the critical values for the various error rates.

For each sample the REXX executable file wrote a one line file, Newseed Data, with an eight digit seed number. The seed number was read by RANDOM SAS which generated a random number and wrote out a data file which set the parameters for the FORTRAN program. These parameters included the random seed number, the number of decimal points requested, the desired variance-covariance matrix, the vector of desired means, the number of variables and the number of observations, and the designated output files. MORRIS EXEC ran the FORTRAN program which generated a listing of the data; this file was read by TRANSFOR SAS which transformed the data into SAS IML matrix form in a flat file. The APDIST SAS program which ran the Andrews-Pregibon statistic on the matrix was edited using an executable file which substituted the new data matrix during each run through the programs. Then APDIST SAS was run; it deleted one observation at a time from the

data matrix, calculated the Andrews-Pregibon statistic, and produced a listing of the statistics for the matrix. A final SAS program, READAP, read the output and appended the data to a data file; therefore, after the 10,438 runs through the REXX exec there was one file containing 1,565,700 Andrews-Pregibon ratios. A simple SAS program, FREQS, ran a frequency distribution on the data file and produced the information needed to do the probability distribution. The programs, with the exception of the FORTRAN program which is quite long, are listed in the Appendices.

Once the frequency distribution was obtained, critical values were identified for alphas of .01, .05, and .10. The values are shown in the table below.

Table of Critical Values for the Andrews-Pregibon Statistic			
p = 3	α = .01	α = .05	α = .10
n = 150	.9260	.9484	.9583

By changing the parameters used in the FORTRAN program, the probability distribution could be developed for other values of n and p. Also, this particular distribution is valid only for data that are multivariate normal (0,1); however, the FORTRAN program can be used to generate data that are from different populations.

Appendix A
REXX EXEC

```
/*    */
ADDRESS COMMAND
SEED = 12345678
TIMES = 1
DO WHILE TIMES < 10,000
SEED = SEED + 99999
"EXECIO 1 DISKW" NEWSEED DATA A "(STRING" SEED
"EXEC SAS RANDOM"
"EXEC MORRIS"
"EXEC SAS TRANSFOR"
"EXEC INMATRX2"
"EXEC SAS APDIST"
"EXEC SAS READAP"
"ERASE APDIST LISTING A"
"ERASE NEWSEED DATA A"
TIMES = TIMES + 1
END
```

Appendix B
RANDOM SAS

```
CMS FILEDEF INDATA DISK NEWSEED DATA A;
CMS FILEDEF OUTDATA DISK DISS DATA A;
/* GENERATE A RANDOM NUMBER */
DATA RANDOM;
  LENGTH X 8;
  INFILE INDATA;
  INPUT SEED1 1-8;
  RETAIN SEED1 2;
  N = 2000;
  P = .2;
  DO I = 1;
    CALL RANBIN (SEED1,N,P,X);
    OUTPUT;
  END;
RUN;
/* REPLACE 8 DIGIT RANDOM NUMBER IN DISS DATA A */
DATA OUTDATA;
  SET RANDOM;
  X = 10000000 + X;
  FILE OUTDATA;
  PUT  @2 '(4F4.3)'/
       @1 '00003 0150    0    1    3   60    9'/
       @1 X/
       @1 '.000.000.000'/
       @1 '1.00.000.000'/
       @1 '.0001.00.000'/
       @1 '.000.0001.00';
```

Appendix C
MORRIS EXEC

```
FILEDEF 50 DISK DISS DATA A (LRECL 80
FILEDEF 60 DISK SCORES LISTING A (LRECL 133
FILEDEF 3 DISK OUTTHO LISTING A (LRECL 133
EXEC ENVSET F
GLOBAL LOADLIB VSF2VECT VSF2LOAD
LOAD MORRIS (CLEAR START
```

Appendix D
TRANSFOR SAS

```
/* TRANSFORMS FORTRAN OUTPUT INTO A MATRIX FOR SAS IML */
CMS FILEDEF INDAT1 DISK SCORES LISTING A (LRECL 133;
CMS FILEDEF OUTDAT1 DISK MATRIX DATA A (LRECL 80 BLKSIZE 80 RECFM
FBS;
DATA RAWDATA;
  INFILE INDAT1 FIRSTOBS = 3;
  INPUT X1 $ 10-16 X2 $ 19-25 X3 $ 28-34;
DATA OUTDAT1;
  SET RAWDATA;
  SEMI = ';';
  FILE OUTDAT1;
    IF _N_ = 1 THEN PUT 'A = {' @;
    ELSE PUT '     '@;
    IF _N_ = 150 THEN PUT X1 ' ' X2 ' ' X3 ' ' '}'SEMI;
    ELSE PUT X1 ' ' X2 ' ' X3 ',';
```

Appendix E
INMATRIX EXEC & INMATRIX XEDIT

INMATRIX EXEC

```
/*    */
'XEDIT APDIST SAS A (PROFILE INMATRIX'
```

INMATRIX XEDIT

```
/*  */
'DOWN 3'
'DELETE 150'
'UP 1'
'GET MATRIX DATA A'
'FILE'
```

Appendix F
APDIST SAS

```
OPTIONS NOCENTER;
PROC IML;
A = {1.9398  0.1359  -1.1587 ,
    -1.2776  0.1060  -0.1976 ,
    -0.0435  0.4709  0.1410 ,
                . . .                ,
                . . .                ,
    0.4692  -0.8134  -1.4504 ,
    0.5467  -1.1297  -0.4473 };
C=A(|,1|);
AT = A';
ATA = (AT)*A;
DETER_A = DET(ATA);
 B = A(|2:150,1:3|);
 BT = B';
 BTB = (BT)*B;
DETER_B = DET(BTB);
AP = DETER_B/DETER_A;
C(|1,|) = AP;
DO I=1 TO 148;
D=A(|1:I,|);
E=A(|(I+2):150,|);
B=D//E;
 BT = B';
 BTB = (BT)*B;
DETER_B = DET(BTB);
AP = DETER_B/DETER_A;
C(|(I+1)|) = AP;
END;
 B = A(|1:149,1:3|);
 BT = B';
 BTB = (BT)*B;
DETER_B = DET(BTB);
AP = DETER_B/DETER_A;
C(|150,|) = AP;
RESET AUTONAME NOPRINT;
PRINT C;
```

Appendix G
READAP SAS

```
/* READS ANDREWS-PREGIBON OUTPUT, PRINTS VALUE */
CMS FILEDEF INDATA DISK APDIST LISTING A;
CMS FILEDEF APDIST DISK APDIST DATA A (LRECL 30 BLKSIZE 30 RECFM FBS;
DATA SDS.TEMP;
  INFILE INDATA;
  FILE APDIST MOD;
  INPUT A $ 1-4 C 9-17;
  IF A = 'ROW';
  PUT C 10.4 @1;
```

Appendix H
FREQS SAS

```
CMS FILEDEF INDATA DISK APDIST DATA A;
DATA DIST;
  INFILE INDATA;
  INPUT AP 1-10;
PROC FREQ;
  TABLES AP;
```

References

Andrews, D. F. & Pregibon, D. (1978). Finding the outliers that matter. Journal of the Royal Statistical Society B, 1, 85-93.

Barnett, V., & Lewis, L. (1978). Outliers in statistical data. Chichester: John Wiley & Sons.

Chatterjee, S. & Hadi, A. S. (1988). Sensitivity analysis in linear regression. New York: John Wiley & Sons.

Fung, W. K. (1990). The Identification of multiple influential observations and outliers in regression. Unpublished manuscript. University of Hong Kong.

Hawkins, D. M., Bradu, D. & Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. Technometrics, 26, 197-208.

Morris, J. D. (1975). A computer program to create a population with any desired centroid and covariance matrix. Educational and Psychological Measurement, 35, 707-710.

Wood, F. S. (1983). Measurements of observations far-out in influence and/or factor space. Paper presented at the Econometrics and Statistics Colloquium at the Chicago Graduate School of Business, Chicago, IL.